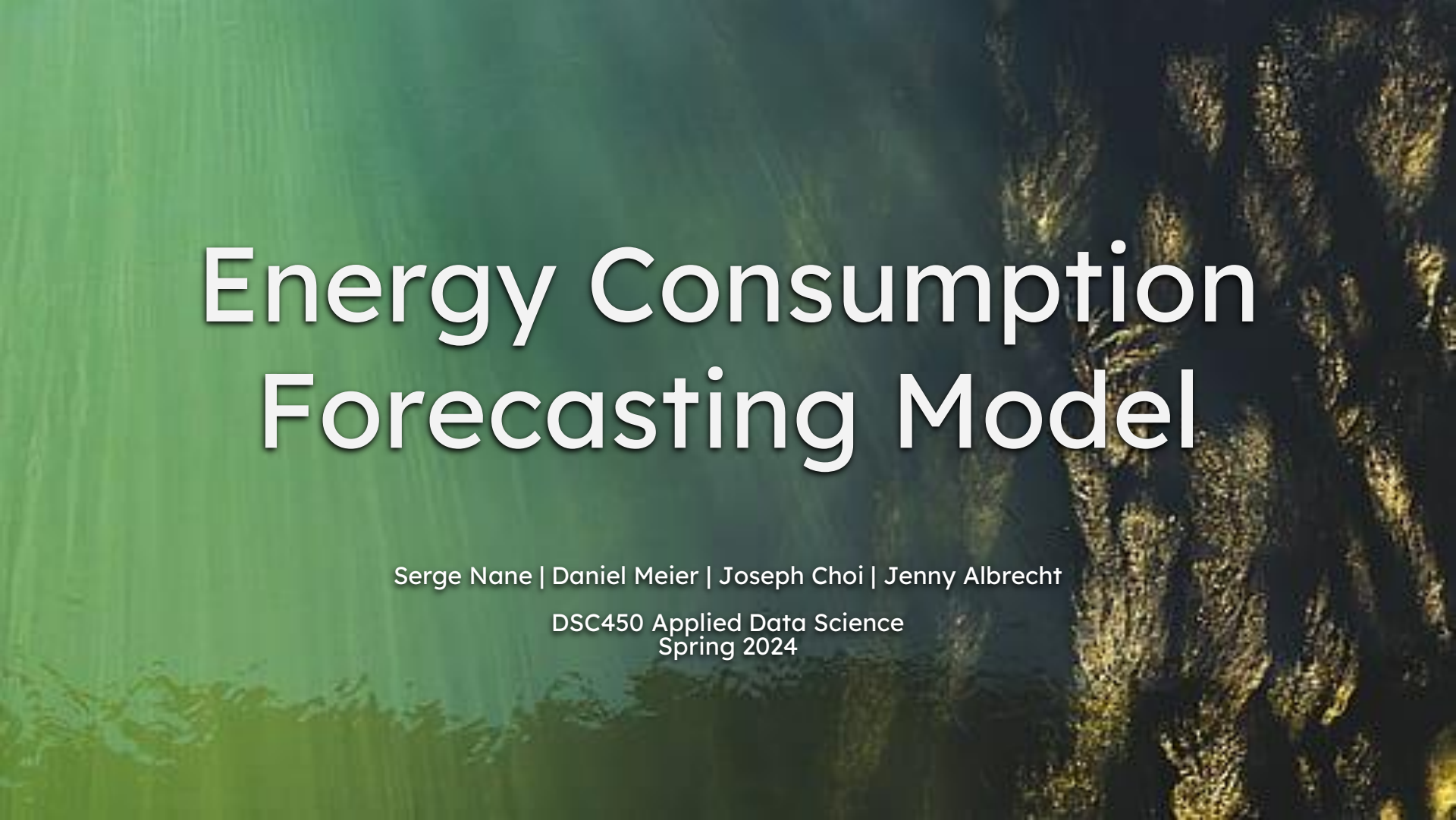


Energy Consumption Forecasting Model

Serge Nane | Daniel Meier | Joseph Choi | Jenny Albrecht

DSC450 Applied Data Science
Spring 2024

AGENDA

INTRODUCTION

- Overview
- Problem statement
- Overview of data
- Method (s) used

MODEL SUMMARY

DATA VISUALIZATION

- Feature importance and Correlation matrix
- Model performance

CONCLUSION

INTRODUCTION

OVERVIEW

- Energy sustainability is crucial for our planet's health and the quality of life of future generations.
- Reducing home energy usage is an essential step towards a more sustainable future.
- Our project uses data analysis and data science to address energy conservation.

PROBLEM STATEMENT

The project aims to answer two key research questions:

1. What are the main factors influencing energy consumption?

2. Can we predict future energy consumption in buildings using historical data?

OVERVIEW OF DATA

- **Description:** Simulated historical energy consumption data, with variables influencing energy usage.
- **Dataset Location:**
  Available on Kaggle Energy-consumption-prediction (kaggle.com)
- **Data Field:** Timestamp, Temperature, Humidity, Square Footage, Occupancy, HVAC Usage, Lighting Usage, Renewable Energy, Day Of Week, Holiday and Energy Consumption

METHOD(S) USED

Approach Overview:

Our strategy leverages machine learning to gain insights into energy efficiency and sustainability from historical data analysis. This includes data visualization for trend identification and algorithm exploration for predictive performance evaluation.

Phase:

- Data Cleaning
- Data Transformation
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Data Modeling

MODEL SUMMARY

Random Forest Model Iterations:

- We conducted seven iterations, varying feature selection and model parameters.
- Initial model (all features, 80/20 split): R2 Score = 0.5417.
- Adjusted to 90/10 split: R2 Score = 0.5799.
- Hyper-tuned parameters: R2 Score = 0.6023.
- Removed two lowest features (AM/PM, Holiday), re-ran: R2 Score = 0.5974.
- Hyper-tuned with updated features: R2 Score = 0.5878.
- Top 3 features only (Temp_F, HVAC Usage, Renewable Energy), base parameters: R2 Score = 0.5370.
- Hyper-tuned with top 3 features: R2 Score = 0.5612.

Linear Regression Model Iterations:

- Conducted five iterations with various feature combinations.
- Initial model (all features, 80/20 split): R2 Score = 0.5987.
- Adjusted to 90/10 split: R2 Score = 0.6224.
- Removed bottom two features: R2 Score = 0.6261.
- Top 3 features only: R2 Score = 0.5693.
- Best 4-feature combination (Temp_F, HVAC Usage, Occupancy, Lighting Usage): R2 Score = 0.6480.

Performance Summary:

- **Random Forest Model:** Achieved moderate accuracy, best R2 Score at 0.6023. Performance varied with feature selection and tuning.
- **Linear Regression Model:** More consistent and higher predictive accuracy, best R2 Score at 0.6480 with a specific feature combination.

Key Predictors:

- Temperature (Temp_F) and HVAC usage emerged as critical factors influencing energy consumption across both models.

Assessment of Accuracy:

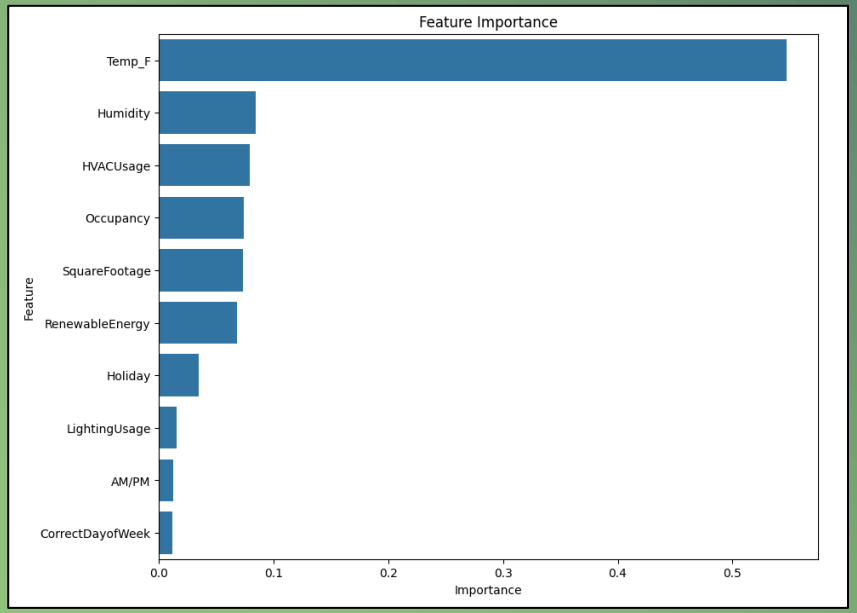
- R2 Scores (0.60 - 0.65 range) indicate reasonable accuracy but not precision. Models can predict trends but may not fully capture dataset complexities.

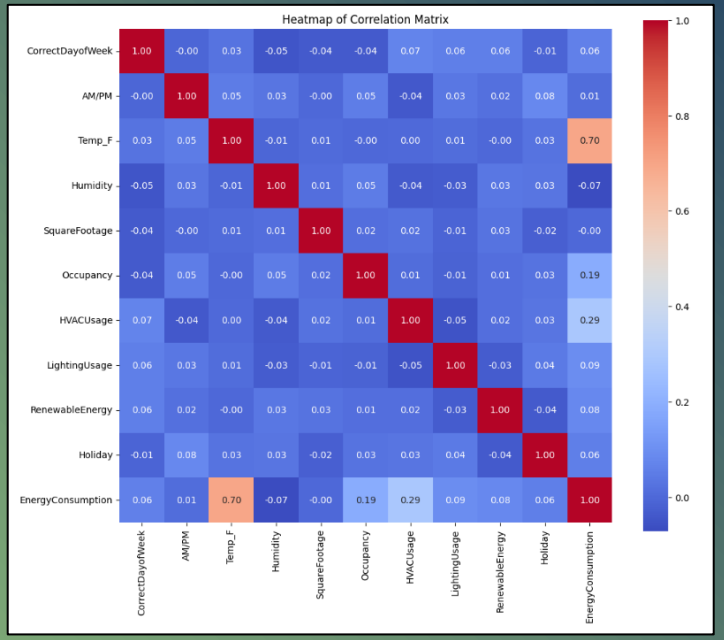
Insights and Implications:

- The improvement in the Linear Regression model with feature optimization suggests a more linear relationship in the data.
- The Random Forest model's varied performance indicates potential non-linear interactions in the data.

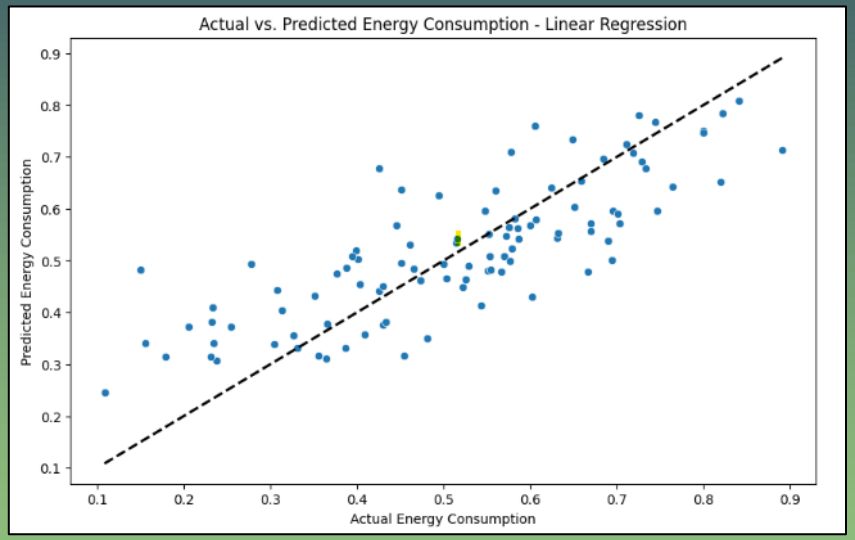
DATA VISUALIZATION

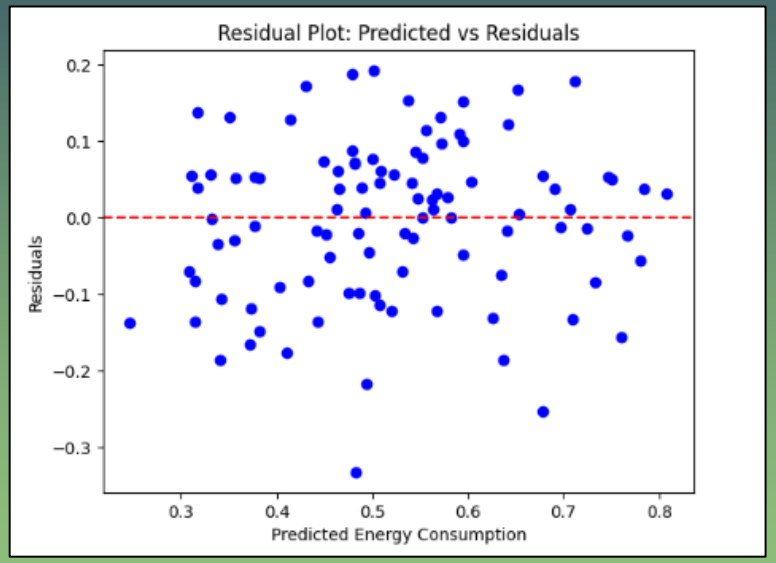
FEATURE IMPORTANCE & CORRELATION MATRIX





Model Performance





CONCLUSION

RECOMMENDATIONS AND NEXT STEPS

Recommendations:

- Install programmable or smart thermostats/HVAC units
- Utilize our energy consumption model as a predictive engine

Next Steps:

- Conduct further analysis to identify additional relevant features
- Perform more feature engineering
- Experiment with different model types and hyperparameters

THANK YOU!